В. В. Нешитой

МОДЕЛИРОВАНИЕ И АНАЛИЗ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ АБИТУРИЕНТОВ ПО РЕЗУЛЬТАТАМ ЦЕНТРАЛИЗОВАННОГО ТЕСТИРОВАНИЯ

Рассмотрим результаты централизованного тестирования (ЦТ) по математике на примере Академии МВД. На ее сайте http://academy.mia.by/ приведены статистические данные по всем абитуриентам, принявшим участие в тестировании (сведения за 2010 г. – *В. Н.*). Их оказалось 597 человек. Баллы варьируются от 1 до 88. Средний балл составил 19,41374. Отсюда следует, что все абитуриенты набрали 11 590 баллов.

Упорядочим абитуриентов по убыванию (точнее, по невозрастанию) баллов, т. е. составим ранговое распределение. Порядковый номер абитуриента будем считать его рангом. Введем обозначения: r – ранг абитуриента; m_r – количество баллов абитуриента с рангом r; $M = \sum_{r=1}^{597} m_r$ – общее количество баллов всех абитуриентов; $p_r = \frac{m_r}{M}$ – доля баллов от общего их количества, приходящаяся на одного абитуриента с рангом r (является оценкой вероятности).

Чтобы извлечь из ранжированного ряда полезную информацию, необходимо рассмотреть некоторые формы представления рангового распределения. Если построить график зависимости $p_r = \Psi(r)$, то получим убывающую кривую, которая быстро приближается к горизонтальной оси. Это значит, что такая форма представления ранговых распределений не дает новой информации.

Часто ранговые распределения графически изображают в двойном логарифмическом масштабе: $\ln p_r = \varphi(\ln r)$. Получается некоторая кривая, которая также несет мало информации.

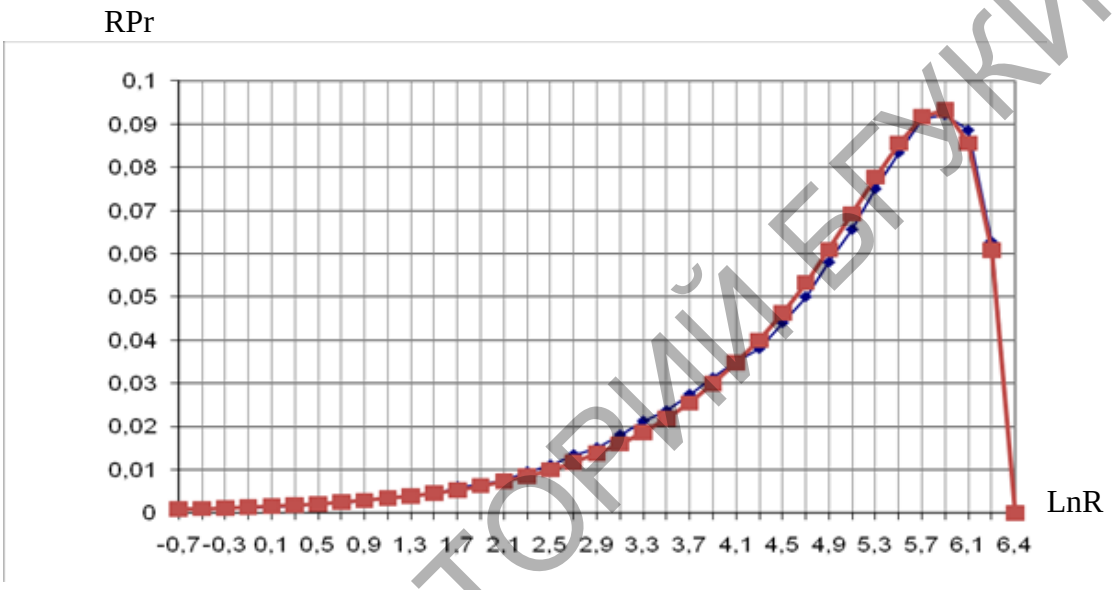
Для извлечения наиболее полной информации из рангового распределения автором предложена другая форма их представления, а именно, в виде зависимости

$$rp_r = f(\ln r). \qquad (1)$$

РЕПОЗИТОРИЙ БГУИР

График зависимости (1) изображает кривую распределения, которая имеет обычно три характерные точки – моду ($\ln r_C$) и две точки перегиба ($\ln r_A$ и $\ln r_B$), причем эти точки располагаются на равных расстояниях от моды [1; 2].

В случае ранговых распределений абитуриентов получается асимметричная кривая с длинной левой ветвью и короткой правой, круто падающей на горизонтальную ось. Точка перегиба *B* в данном случае отсутствует (см. рисунок, на котором эмпирическая кривая отмечена мелкими квадратами).



Кривая рангового распределения абитуриентов
в системе координат (LnR; RPr)

Для дальнейшего анализа статистической кривой распределения необходимо вычислить теоретический закон. Ранговые распределения могут быть описаны второй системой непрерывных распределений автора [1; 2]

$$p(t) = Nt^{k\beta-1}\left(-\alpha u t^{\beta}\right)^{\frac{1}{u}-1}. \qquad (2)$$

Здесь *p(t)* – плотность распределения; *N* – нормирующий множитель; α, β, k, u – параметры, которые для каждого статистического распределения принимают свои конкретные значения и вычисляются по статистическим данным.

Эта плотность при определенных значениях параметров может быть убывающей. Если ее преобразовать к виду $tp(t) = f(\ln t)$, то теоретическая кривая распределения будет иметь моду С и две

РЕПОЗИТОРИЙ БГУКИ

точки перегиба *A* и *B* (при u<0,5). Примем эти точки в качестве границ ядра абитуриентов и зон рассеяния. Между этими точками существует соотношение [1; 2]

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n,$$

которое по праву может считаться математически точной формулировкой закона рассеяния информации в смысле Бредфорда. Здесь величина *n* вычисляется по формуле

$$n = \left[1 + \frac{1 - u + \sqrt{k(1+ku-u) + (1-u)(1-u)}}{2k(1+ku-u)}\right]^{\frac{1}{\beta}}.$$

Абсциссы точек перегиба в системе координат (*lnt*; *tp(t)*) вычисляются по формулам

$$t_A = t_C / n;\ t_B = t_C \cdot n,$$

где $t_C = \left(\frac{k}{\alpha(1+ku-u)}\right)^{\frac{1}{\beta}}$.

Вычисление рангового закона распределения

Для вычисления четырехпараметрического закона распределения и оценок его параметров классические методы (метод моментов, метод максимального правдоподобия) оказались непригодными. Поэтому автором теории обобщенных распределений были разработаны свои методы: универсальный метод моментов и общий устойчивый метод, которые применимы ко всем трем системам непрерывных распределений. При разработке этих методов было принято решение отказаться от выдвижения гипотез, а также от совместного решения четырех уравнений с четырьмя неизвестными (по числу параметров обобщенных распределений).

Задача установления закона распределения и вычисления оценок его параметров разбивается на два этапа. На первом этапе по статистическому распределению вычисляются два показателя (*B* – асимметрии и *H* – островершинности), которые зависят от двух параметров формы *k, u.* По этим показателям с помощью заранее построенной бинарной сетки (номограммы) устанавливается тип теоретического распределения и находятся оценки параметров *k, u* (либо графически, либо теоретически путем решения системы двух уравнений с двумя неизвестными). Оценки остальных параметров

РЕПОЗИТОРИЙ БГУКИ

α, β вычисляются по относительно простым формулам.

При использовании устойчивого метода оценивания параметров обобщенной плотности (2) для вычисления статистических показателей используются формулы

$$\left.\begin{aligned} &\nu_1^* = \overline{\ln t} = \sum_{i=1}^{n} \ln t_i \frac{m_i}{M};\ S_1^* = \sum_{i=1}^{n} t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} \\ &S_3^* = \sum_{i=1}^{n} t_i^3 \left(\frac{m_i}{M}\right)^4 \frac{1}{h^3};\ H^* = \frac{S_3^*}{(S_1^*)^3} \\ &B^* = \sum_{i=1}^{n} t_i \ln t_i \left(\frac{m_i}{M}\right)^2 \frac{1}{h_i} - \nu_1^* S_1^* \end{aligned}\right\}$$

По показателям *B, H* с помощью номограммы [1; 2] устанавливается тип аппроксимирующего распределения и находятся оценки параметров *k, u* (по крайней мере в первом приближении).

Далее вычисляются оценки параметров β, α (или произведение αu)

$\beta = \frac{S_1}{S_1^{(z)}}$; $\alpha u = e^{(\nu_1^{(z)} - \beta \nu_1)}$ – для распределений первого типа.

Величины $S_1^{(z)}$, $\nu_1^{(z)}$ рассчитываются теоретически при известных оценках параметров *k, u*. Например, в случае распределения первого типа имеем

$$S_1^{(z)} = \frac{1}{2\sqrt{\pi}} \frac{2\left(k+\frac{1}{u}\right)-1}{\frac{2}{u}-1} \cdot \frac{g(k)g\left(\frac{1}{u}\right)}{g\left(k+\frac{1}{u}\right)},\quad \nu_1^{(z)} = \pm\left[\Psi(k) - \Psi\left(k+\frac{1}{u}\right)\right],$$

при этом величины $g(x)$ и $\Psi(x)$ вычисляются по формулам

$$g(x) \approx \sqrt{x} e^{-\frac{1}{8x}+\frac{1}{192x^3}-\frac{1}{640x^5}+\ldots},\quad \psi(x) \approx -\left(\frac{1}{x}+\frac{1}{x+1}\right) + \ln(x+2) - \frac{1}{2(x+2)} - \frac{1}{12(x+2)^2} + \ldots$$

При 0<x≤4 значения этих величин могут быть взяты из таблицы [1].

В результате расчетов по программе автора SNR1V97 были получены следующие результаты: B = 0,160983; H = 1,606391; аппроксимирующее распределение относится к первому типу с параметрами: αu = 4,293446E–05; ß = 1,550406; k = 0,506847; u = 0,540161. Нормирующий множитель N = 6,930736E–03.

РЕПОЗИТОРИЙ БГУКИ

Теоретическое распределение (см. рис.), представленное в форме $tp = f(\ln t)$, имеет моду $\ln t_C$ = 5,850189 (t_C = 347) и одну точку перегиба $\ln t_A$ = 5,165518 (t_A=175). При этом накопленная доля баллов до точек А и С соответственно равна:

F (t_A) = 0,492 – по расчету; 0,498 – по факту.

F (t_C) = 0,789 – по расчету; 0,781 – по факту.

Таким образом, 175 абитуриентов с наиболее высокими баллами (от 22 до 88) образуют ядро, на которое приходится около половины всех баллов.

Далее идет первая зона рассеяния (от точки А до точки С). Ее образуют $t_C - t_A$ = 347 – 175 = 172 абитуриента. На первую зону рассеяния приходится 28,3% баллов (78,1–49,8=28,3) – по статистическому ранговому распределению или 28,8% – по теоретическому распределению. Сюда вошли абитуриенты с баллами от 16 до 22.

Остальные абитуриенты составляют вторую зону рассеяния. Их число равно: 597 – 347 = 250. На вторую зону рассеяния приходится 22% баллов. Количество баллов в этой зоне колеблется от 1 до 15 .

Рассмотрим далее соотношения между долями баллов и долями абитуриентов в каждой зоне (см. табл.).

Зона	Доля баллов, %	Доля абитуриентов, %	Доля баллов/доля абитуриентов
Ядро	49,8	29,3	1,70
I	28,3	28,8	0,98
II	21,9	41,9	0,52

Из анализа приведенной таблицы, а также графика можно сделать однозначный вывод: из второй зоны рассеяния зачислять абитуриентов в вузы нельзя. В рассмотренном случае они составили 42% от общего числа тестируемых, но набрали лишь 22% баллов.

С помощью ранговых распределений можно проверить эффективность тестирования. Для этого достаточно провести тестирование школьников 4–5 классов по программе ЦТ. В результате должны получиться данные, близкие к случайным ответам. Сравнительный анализ двух ранговых распределений

РЕПОЗИТОРИЙ БГУКИ

покажет, насколько результаты тестирования абитуриентов близки к случайным ответам. Об этой близости уже свидетельствует низкое среднее значение баллов – 19,4. Другой метод – расчет баллов при случайных ответах и сравнение их с фактическими баллами.

---

1. Нешитой, В. В. Методы статанализа в библиотечной деятельности: вычисление непрерывных распределений: учеб.-метод. пособие / В. В. Нешитой. – Минск: Белорус. гос. ун-т культуры и искусств, 2010. – 61 с.

2. Нешитой, В. В. Элементы теории обобщенных распределений: монография / В. В. Нешитой. – Минск: РИВШ, 2009. – 204 с.

РЕПОЗИТОРИЙ БГУКИ